

Аннотация дипломной работы

Тема: «Разработка высоконагруженного сервиса идентификации контентных дублей»

Исполнитель: Грицкевич Евгений Марьянович

Руководитель: Пазюра Екатерина Васильевна, старший преподаватель кафедры ТП

Кафедра: кафедра технологий программирования, специальность «Прикладная информатика», специализация «Программное обеспечение встроенных систем».

Объем работы: Дипломная работа, 48 с., 16 рис., 5 источников

Ключевые слова: ВЕБ-СЕРВИС, КОНТЕНТНЫЕ ДУБЛИ, ВЫСОКАЯ ПРОИЗВОДИТЕЛЬНОСТЬ

В дипломной работе рассматривается задача разработки сервиса поиска контентных дублей текстовых документов. К сервису предъявляются высокие требования к скорости ответа на запрос пользователя при большом количестве документов подлежащих сравнению. Целью работы является – разработка и тестирование сервиса, выполняющего поиска контентных дублей документов. В работе описана предметная область поиска нечетких дубликатов документов и проанализированы алгоритмы для быстрого поиска «похожих» документов. Данные алгоритмы использованы для построения двухуровневой схемы поиска дубликатов. Проведены эксперименты для определения скорости работы полученных алгоритмов в разработанном программном средстве и сравнение результатов. Разработанный сервис может быть использован для поиска дублей в большой обновляющейся коллекции документов за минимальное время.